We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



21

Recognition of Paralinguistic Information using Prosodic Features Related to Intonation and Voice Quality

Carlos T. Ishi
ATR
Japan

1. Introduction

Besides the linguistic (verbal) information conveyed by speech, the paralinguistic (non-verbal) information, such as intenning the classification of paralinguistic information. Among the several paralinguistic items extions, attitudes and emotions expressed by the speaker, also convey important meanings in communication. Therefore, to realize a smooth communication between humans and spoken dialogue systems (such as robots), it becomes important to consider both linguistic and paralinguistic information.

There is a lot of past research concerpressing intentions, attitudes and emotions, most previous research has focused on the classification of the basic emotions, such as anger, happiness and sadness (e.g., Fernandez et al., 2005; Schuller et al., 2005; Nwe et al., 2003; Neiberg et al., 2006). Other works deal with the identification of attitudes and intentions of the speaker. For example, Fujie et al. (2003) report about the identification of positive/negative attitudes of the speaker, while Maekawa (2000) reports about the classification of paralinguistic items like admiration, suspicion, disappointment and indifference. In Hayashi (1999), paralinguistic items like affirmation, asking again, doubt and hesitation were also considered. In the present work, aiming at smooth communication in dialogue between humans and spoken dialogue systems, we consider a variety of paralinguistic information, including intentions, attitudes and emotions, rather than limiting our focus to the basic emotions.

The understanding of paralinguistic information becomes as important as linguistic information in spoken dialogue systems, especially in interjections such as “eh”, “ah”, and “un”. Such interjections are frequently used to express a reaction to the conversation partner in a dialogue scenario in Japanese, conveying some information about the speaker’s intention, attitude, or emotion. As there is little phonetic information represented by such interjections, most of the paralinguistic information is thought to be conveyed by its speaking style, which can be described by variations in prosodic features, including voice quality features.

So far, most previous research dealing with paralinguistic information extraction has focused only on intonation-related prosodic features, using fundamental frequency (F0), power and duration (e.g., Fujie et al., 2003; Hayashi, 1999). Others also consider segmental features like cepstral coefficients (e.g., Schuller et al., 2005; Nwe et al., 2003). However,

Open Access Database www.intechweb.org

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

analyses of natural conversational speech have shown the importance of several voice quality features caused by non-modal phonations (e.g., Klasmeyer et al., 2000; Kasuya et al., 2000; Gobl et al., 2003; Campbell et al., 2003; Fujimoto et al., 2003; Erickson, 2005).

The term "voice quality" can be used in a broad sense, as the characteristic auditory colouring of an individual speaker's voice, including qualities such as nasalized, dentalized, and velarized, as well as those brought about by changing the vocal tract length or hypopharyngeal area (e.g., Imagawa et al., 2003; Kitamura et al., 2005; Dang et al., 1996). Here, we use it in a narrow sense of the quality deriving solely from laryngeal activity, i.e., from different vibration modes of the vocal folds (different phonation types), such as breathy, whispery, creaky and harsh voices (Laver, 1980).

Such non-modal voice qualities are often observed especially in expressive speech utterances, and should be considered besides the classical intonation-related prosodic features. For example, whispery and breathy voices are characterized by the perception of a turbulent noise (aspiration noise) due to air escape at the glottis, and are reported to correlate with the perception of fear (Klasmeyer et al., 2000), sadness, relaxation and intimateness in English (Gobl et al., 2003), and with disappointment (Kasuya et al., 2000; Fujimoto et al., 2003) or politeness in Japanese (Ito, 2004). Vocal fry or creaky voices are characterized by the perception of very low fundamental frequencies, where individual glottal pulses can be heard, or by a rough quality caused by an alternation in amplitude, duration or shape of successive glottal pulses. Vocal fry may appear in low tension voices correlating with sad, bored or relaxed voices (Klasmeyer et al., 2000; Gobl et al., 2003), or in pressed voices expressing attitudes/feelings of admiration or suffering (Sadanobu, 2004). Harsh and ventricular voices are characterized by the perception of an unpleasant, rasping sound, caused by irregularities in the vocal fold vibrations in higher fundamental frequencies, and are reported to correlate with anger, happiness and stress (Klasmeyer et al., 2000; Gobl et al., 2003).

Further, in segments uttered by such voice qualities (caused by non-modal phonation types), F0 information is often missed by F0 extraction algorithms due to the irregular characteristics of the vocal fold vibrations (Hess, 1983). Therefore, in such segments, the use of only prosodic features related to F0, power and duration, would not be enough for their complete characterization. Thus, other acoustic features related to voice quality become important for a more suitable characterization of their speaking style.

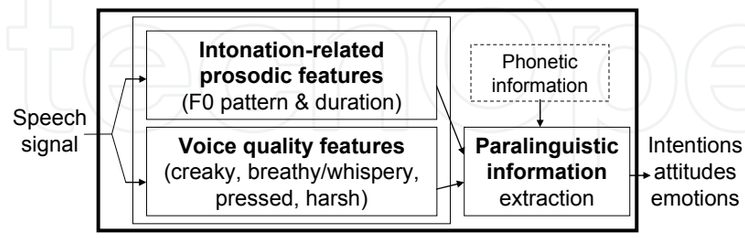


Fig. 1. Block diagram of the proposed framework for paralinguistic information extraction.

Fig. 1 shows our framework proposed for extraction of paralinguistic information, by using information of voice quality features, in addition to intonation-related prosodic features. In our previous research, we have proposed several acoustic parameters for representing the features of intonation and specific voice qualities (Ishi, 2004; Ishi, 2005; Ishi et al., 2005). In the present chapter, evaluation on the performance of the acoustic parameters in the automatic recognition of paralinguistic information is presented.

The rest of the chapter is organized as follows. In Section 2, the speech data and the perceptual labels used in the analysis are described. Section 3 describes the acoustic parameters representing prosodic and voice quality features. In Section 4, the automatic detection of paralinguistic information is evaluated by using the acoustic parameters described in Section 3, and Section 5 concludes the chapter.

2. Description of the speech data and the perceptual data

2.1 Description of the speech data for analysis and experimental setup

In the present research, the Japanese interjections "e" and "un" (including variations such as "e", "eh", "ee", "eeee" and "un", "uun", "hun", "nnn", "uhn", etc.) are chosen for analysis and evaluation. These interjections are often used to express a reaction in Japanese conversational speech, and convey a large variety of paralinguistic information depending on its speaking style (Campbell et al., 2004). Possible paralinguistic information (intentions, attitudes or emotions) transmitted by varying the speaking styles of the interjections "e" and "un" are listed in Table 1.

Original adjectives and translations	PI item
"koutei", "shoudaku" (affirm, accept)	***Affirm***
"dooi", "rikai", "nattoku" (agree, understand, consent)	***Agree***
"aiduchi" (backchannel: agreeable responses)	***Backchannel***
"hitei" (deny, negate)	***Deny***
"kangaechuu", "filler" (thinking, filler)	***Thinking***
"mayoi", "konwaku", "tomadoi", "nayamu", "chuucho" (embarrassed, undecided, hesitated)	***Embarrassed***
"kikikaeshi" (ask for repetition)	***AskRepetition***
"bikkuri", "odoroki" (surprised, amazed, astonished)	***Surprised***
"igai" (unexpected)	***Unexpected***
"utagai", "gimon" (suspicious, doubt)	***Suspicious***
"hinan", "kyozetsu" (blame, criticise, reject)	***Blame***
"ken'o", "iya" (disgusted, disliked)	***Disgusted***
"fuman" (dissatisfied, frustrated)	***Dissatisfied***
"kanshin" (admired)	***Admired***
"senbou", "urayamashii" (envious)	***Envious***

Table 1. List of paralinguistic information conveyed by the interjections "e" and "un".

The list of Table 1 was obtained by referring to a list of speech acts annotated for the interjections "e" and "un" in the JST/CREST ESP conversational speech database (JST/CREST ESP Project homepage). The items of the list have been obtained by free-text annotations of four subjects, in "e" and "un" utterances appearing in natural conversations. The annotated words have been arranged by the four subjects for reducing redundancies. We do not guarantee that this list contains all the possible paralinguistic information that the interjections "e" and "un" can convey. However, we consider that this list is rich enough for our purposes of human-machine communication.

The list of Table 1 includes paralinguistic items expressing some intention, such as affirm and ask for repetition, some attitude, such as suspicious and dissatisfied, and some emotion, such as surprised and disgusted. These items are more related to intentions or speech acts

conveyed by the utterances, rather than the basic emotions, such as anger, happy and sadness. Since it is difficult to clearly classify these items as intentions, attitudes or emotions, in the present research we simply call them paralinguistic information (PI) items.

In the present research, speech data was newly recorded in order to get a balanced data in terms of the PI conveyed by the interjections "e"/"un". For that purpose, sentences were elaborated in such a way to induce the subject to produce a specific PI. Some short sentences were also elaborated to be spoken after the interjections "e"/"un", in order to get a reaction as natural as possible. Two sentences were elaborated for each PI item of Table 1, by two native speakers of Japanese. (Part of the sentences is shown in the Appendix.)

The sentences were first read by one native speaker. These sentences will be referred to as "inducing utterances". Then, subjects were asked to produce a target reaction, i.e., utter in a way to express a specific PI, through the interjection "e", after listening to each pre-recorded inducing utterance. The same procedure was conducted for the interjection "un". A short pause was required between "e"/"un" and the following short sentences. Further, the utterance "he" (with the aspirated consonant /h/ before the vowel /e/) was allowed to be uttered, if the subject judged that it was more appropriate for expressing some PI.

Utterances spoken by six subjects (two male and four female speakers between 15 to 35 years old) are used for analysis and evaluation. In addition to the PI list, speakers were also asked to utter "e" and "un" with a pressed voice quality, which frequently occurs in natural expressive speech (Sadanobu, 2004), but was found more difficult to naturally occur in an acted scenario. Of the six speakers, four could produce pressed voice utterances. The data resulted in 173 "e" utterances, and 172 "un" utterances.

For complementing the data in terms of voice quality variations, another dataset including utterances of a natural conversational database (JST/CREST ESP database) was also prepared for evaluation. The dataset is composed of 60 "e" utterances containing non-modal voice qualities, extracted from natural conversations of one female speaker (speaker FAN), resulting in a total of 405 utterances for analysis.

All the "e" and "un" utterances were manually segmented for subsequent analysis and evaluation.

2.2 Perceptual labels of paralinguistic information (PI) items

Perceptual experiments were conducted to verify how good the intended (induced) PI items could be correctly recognized when listening only to the monosyllabic utterances, i.e., in a context-free situation. The purpose is to verify the ambiguity in the expression of a PI item, since the same speaking style could be used to express different PI items, in different contexts.

Three untrained subjects (who are different from the speakers) were asked to select from the PI item list of Table 1, one or multiple items that could be expressed by each stimuli (i.e., the segmented "e"/"un" utterances). Multiple items were allowed since the same speaking style could be used to express different PI items. As a result of the multiple selections, two, three and more than three labels were selected in 40%, 14% and 4% of the utterances, respectively. Regarding the subjects' agreement, in 51% of the utterances, all three subjects agreed in assigning the same PI items, while in 93% of the utterances, at least two of the three subjects agreed in assigning the same PI items.

Fig. 2 shows the matches, mismatches and ambiguities between intended and perceived PI items, when listening only to the "e" utterances, i.e., in a context-free situation. The perceptual degrees (between 0 to 1) are computed by counting the number of labels of a

perceived PI item in all utterances of an intended PI item, and dividing by the number of utterances, and by the number of subjects.

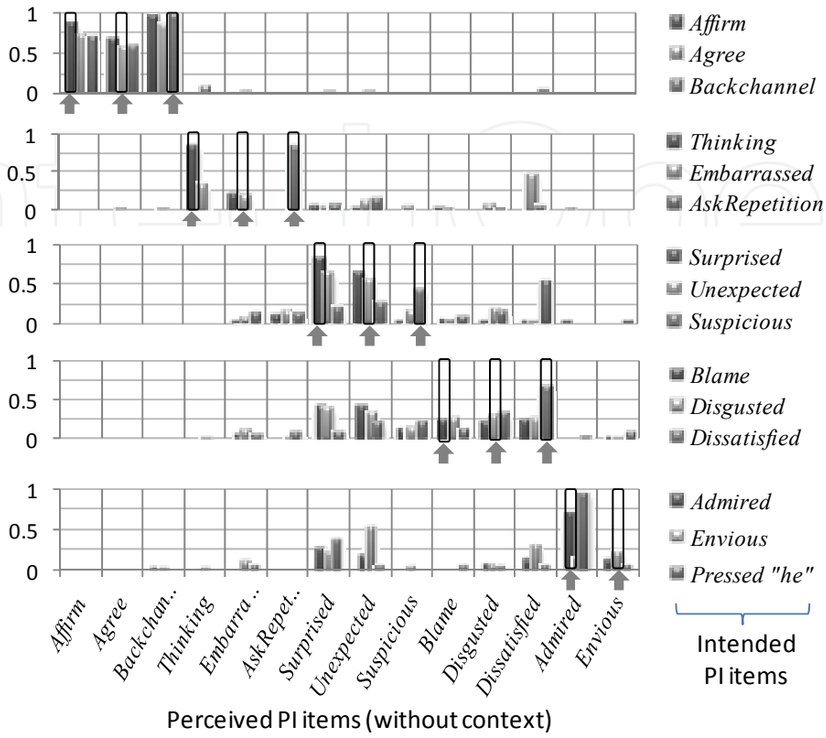


Fig. 2. Perceptual degrees of the intended PI items of "e" utterances (without context, i.e., by listening only to the interjections). The bars indicated by arrows show the matching degrees between intended and perceived items.

First, regarding the matching between intended and perceived PI items, it can be observed in the bars indicated by arrows in Fig. 2 that Affirm, Backchannel, Thinking, AskRepetition and Surprised show high matching degrees, while Agree, Unexpected, Suspicious, Dissatisfied and Admired show moderate matching. However, Embarrassed, Blame, Disgusted, and Envious show very low matching degrees, indicating that the intended PI could not be perceived in most of their utterances, in a context-free situation.

The mismatches and ambiguities between intended and perceived PI items are shown by the bars excluding the ones indicated by arrows in Fig. 2. Among the PI items with large mismatches, most in Embarrassed is perceived as Thinking or Dissatisfied, while most in Unexpected is perceived as Surprised. Some of the confusions are acceptable, since there may be situations where the speaker is thinking while embarrassed, or where the speaker feels surprised and unexpected at the same time. Confusion is also found between samples of Blame, Disgusted, Dissatisfied and Suspicious. This is also an acceptable result, since all these PI items express negative reactions.

However, Surprised, Unexpected and Dissatisfied are perceived in the stimuli of several intended PI items. This indicates that the identification of these PI items would only be possible by considering context information, for example, by taking into account the sentence following the "e" utterances.

Further, even among the PI items where a good matching was achieved between intended and perceived items, ambiguity may exist between some of the PI items. For example, there is high confusion between the stimuli of Affirm, Agree and Backchannel, but no confusion between these and other PI items.
Finally, regarding the pressed voice samples, pressed "he" was mostly perceived as Admired, while pressed "e" (omitted from Fig. 2) was perceived as Embarrassed, Disgusted or Dissatisfied.
The results above imply that an automatic detection of these ambiguous items will also probably be difficult based only on the speaking style of the utterance "e", i.e., without using context information.
From the results above, we can predict that many of the PI items can not be identified without context information. However, we can expect that some groups of PI items can be roughly discriminated, even when context is not considered: {Affirm/Agree/Backchannel}, {Thinking/Embarrassed}, {AskRepetition}, {Surprised/Unexpected}, {Blame/Disgusted/ Dissatisfied/Suspicious}, and {Admired/Envious}. These PI groups will be used as a basis to evaluate how much they can be discriminated by the use of intonation and voice quality-related prosodic features in "e"/"un" utterances (i.e., without context information).
Finally, 35 of the 405 utterances, corresponding to the mismatches between different PI groups, were considered as badly-acted, and were removed from the subsequent evaluation of automatic identification.

2.3 Perceptual voice quality labels and relationship with paralinguistic information

Perceptual voice quality labels are annotated for two purposes. One is to verify their effects in the representation of different PI items. Another is to use them as targets for evaluating the automatic detection of voice qualities.
The perceptual voice quality labels are annotated by one subject with knowledge about laryngeal voice qualities (the first author), by looking at the waveforms and spectrograms, and listening to the samples. Samples for several voice quality labels can be listened in the Voice quality sample homepage. The voice quality labels are annotated according to the following criteria.

- m: modal voice (normal phonation).
- w: whispery or breathy voices (aspiration noise is perceived throughout the utterance).
- a: aspiration noise is perceived in the offset of the last syllable of the utterance.
- h: harsh voice (rasping sound, aperiodic noise) is perceived.
- c: creaky voice or vocal fry is perceived.
- p: pressed voice is perceived.
- Combination of the above categories: for example, hw for harsh whispery, and pc for pressed creaky.

A question mark "?" was added for each voice quality label, if their perception were not clear. Fig. 3 shows the distributions of the perceived voice quality categories for each PI item.
We can first observe in Fig. 4 that soft aspiration noise (w?) is perceived in some utterances of almost all PI items. In contrast, strong aspiration noise (w), harsh or harsh whispery voices (h, hw) and syllable offset aspiration noise (a, a?) are perceived in PI items expressing some emotion or attitude (Surprised, Unexpected, Suspicious, Blame, Disgusted, Dissatisfied and Admired). This indicates that the detection of these voice qualities (w, h,

hw, a) could be useful for the detection of these expressive PI items. The soft aspiration noise (w?) appearing in emotionless items (Affirm, Agree, Backchannel and Thinking) is thought to be associated to politeness (Ito, 2004).

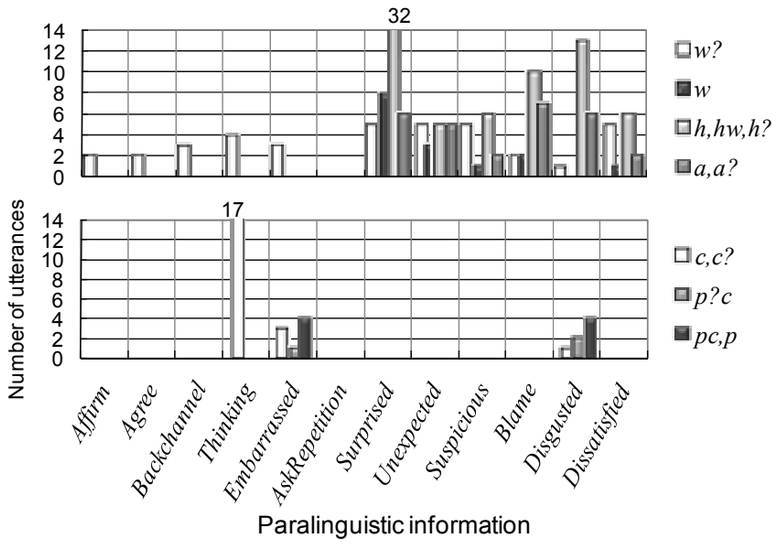


Fig. 3. Distribution of perceived categories of whispery/breathy/aspirated (w, a), harsh (h, hw), creaky (c), and pressed (p) voice qualities, for each paralinguistic information item.

Regarding to creaky voices (c), we can observe in the figure that they are perceived in Thinking, Embarrassed, Disgusted and Admired. However, the additional perception of pressed voices (p) is important to discriminate between emotionless fillers (Thinking), and utterances expressing some emotion or attitude (Admired, Disgusted and Embarrassed).
It is worth mentioning that the use of non-modal voice qualities is not strictly necessary for expressing an emotion or attitude, since different speakers may use different strategies to express a specific PI item. However, the results of the present section imply that when a non-modal voice quality occurs in an utterance, it will probably be associated with an emotion or an attitude.

3. Acoustic parameters representing prosodic features related to F0, duration and voice quality

In this section, we describe acoustic parameters that potentially represent the perception of intonation and voice quality-related prosodic features, which are responsible for the discrimination of different speaking styles, and verify their performance in automatic detection.

3.1 Acoustic parameters related to intonation-related prosodic features: F0move and duration

The main acoustic features used for intonation-related prosodic features are fundamental frequency (F0), power and segmental duration. In the present study, we avoid the use of power as a prosodic feature, due to its large variability caused by the microphone gains, the difference in microphone types, the distance between the mouth and the microphone, and the background noise.

In Ishi (2005), a set of parameters was proposed for describing the intonation of phrase finals (phrase final syllables), based on F0 and duration information. Here, we use a similar set of parameters with some modifications, for the monosyllabic "e" and "un" utterances.
For the pitch-related parameters, the F0 contour is estimated as a first step. In the present research, F0 is estimated by a conventional method based on autocorrelation. Specifically, the normalized autocorrelation function of the LPC inverse-filtered residue of the pre-emphasized speech signal is used. However, any algorithm that can reliably estimate F0 could be used instead. All F0 values are then converted to a musical (log) scale before any subsequent processing. Expression (1) shows a formula to produce F0 in semitone intervals.

$$F0[\text{semitone}] = 12 * \log2 (F0[\text{Hz}]) \quad (1)$$

In the second step, each (monosyllabic) utterance is broken in two segments of equal length, and representative F0 values are extracted for each segment. In Ishi (2005), several candidates for the representative F0 values have been tested, and here, we use the ones that best matched with perceptual scores of the pitch movements. For the first segment, an average value is estimated using F0 values within the segment (F0avg2a). For the second segment, a target value is estimated as the F0 value at the end of the segment of a first order regression line of F0 values within the segment (F0tgt2b). In other words, for the first portion of the utterance, an average or predominant F0 value is perceived, while in the final portion, a target value to where F0 is moving is perceived.
A variable called F0move is then defined as the difference between F0tgt2b and F0avg2a, as shown in expression (2), quantifying the amount and direction of the F0 movement within a syllable.

$$F0move2 = F0tgt2b - F0avg2a \quad (2)$$

F0move is positive for rising F0 movements, and negative for falling movements. It has been shown that F0move parameters match better with the human pitch perception, rather than linear regression-based slope parameters. Details about the evaluation of the correspondence between F0move and perceptual features can be found in Ishi (2005).
The representation of F0 movements by F0move is valid when F0 only rises, only falls, or does not change within a syllable. This condition is true for most cases in Japanese syllables. However, there are cases where F0 falls down and then rises up within the same syllable. For example, a fall-rise intonation is commonly used in "un" utterances for expressing a denial.
In the present research, we proposed a method for detecting fall-rise movements, by searching for negative F0 slopes in the syllable nucleus, and positive F0 slopes in the syllable end portion. Here, syllable nucleus is defined as the 25 % to 75 % center portion of the syllable duration, while the syllable end is defined as the 40 % to 90 % portion of the syllable. The initial and final portions of the syllable are removed from the slope searching procedure, in order to avoid misdetection of F0 movements due to co-articulation effects.
If a fall-rise movement is detected, the syllable is divided in three portions of equal length. The representative F0 value of the first portion is estimated as the average F0 value (F0avg3a). For the second portion, the minimum F0 value (F0min3b) is estimated. Finally, for the last portion, a target value (F0tgt3c) is estimated in the same way of F0tgt2b. Then, two F0move values are estimated according to the expressions

$$F0move3a = F0min3b - F0avg3a, \quad (3)$$

$$F0move3b = F0tgt3c - F0min3b, \quad (4)$$

representing the falling and rising degrees, respectively. Fig. 4 shows a schematic procedure for F0move estimation.

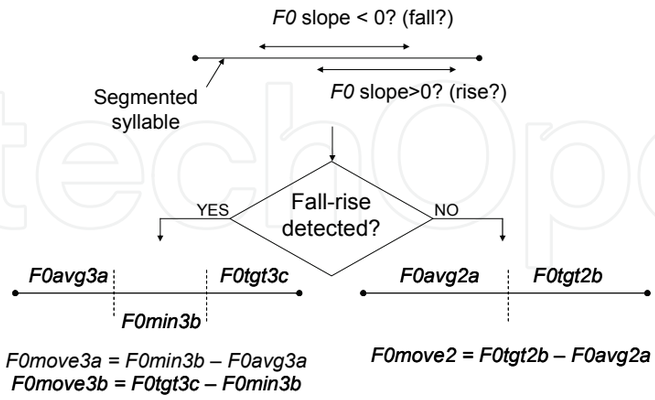


Fig. 4. Schematic procedure for estimation of F0move parameters in the monosyllabic utterances.

Fall-rise tones were correctly detected in all "un" utterances expressing denial. They were also detected in two "e" utterances. However, in these two cases, the F0move values of the falling movement were smaller than 2 semitones, and were not perceived as a fall-rise movement. In contrast, the F0move values for the "un" utterances expressing denial were all larger than 3 semitones, and clearly perceived as fall-rise tones.

For utterance duration, the manually segmented boundaries could be directly used, since the utterances are monosyllabic. However, as the manual segmentation may contain some silence (non-speech) portions close to the segmentation boundaries, an automatic procedure was further conducted, by estimating the maximum power of the syllable, and moving the boundaries until the power becomes 20 dB weaker than the maximum power. The newly segmented boundary intervals were used as segmental duration.

3.2 Detection of vocal fry (creaky voice): PPw, IFP, IPS

Vocal fry or creaky voices are characterized by the perception of very low fundamental frequencies, where individual glottal pulses can be heard, or by a rough quality caused by an alternation in amplitude, duration or shape of successive glottal pulses.

In the present research, we use the algorithm proposed in Ishi et al. (2005) for detection of vocal fry segments. A simplified block diagram of the detection algorithm is shown in Fig. 5. The algorithm first searches for power peaks in a "very short-term" power contour (obtained by using 4 ms frame length each 2 ms), which reflects the impulse-like properties of the glottal pulses in very low fundamental frequencies, characteristic of vocal fry signals. Then, it checks for constraints of periodicity and similarity between successive glottal pulses.

The periodicity constraint is to avoid the misdetection of a modal (periodic) segment between two glottal pulses. The similarity constraint is to avoid the misdetection of impulsive noises, assuming that during speech the vocal tract moves slowly enough so that the shapes of consecutive glottal pulses are similar.

The algorithm depends basically on three parameters.

- PPw : power thresholds for detection of power peaks in the very short-term power contour;
- IFP: intra-frame periodicity, which is based on the normalized autocorrelation function;
- IPS: inter-pulse similarity, which is estimated as a cross-correlation between the speech signals around the detected peaks.

Here, vocal fry segments are detected by using PPw larger than 7 dB, IFP smaller than 0.8, and IPS larger than 0.6. These thresholds are based on the analysis results reported in Ishi et al. (2005). Details about the evaluation of each parameter can be found in Ishi et al. (2005).

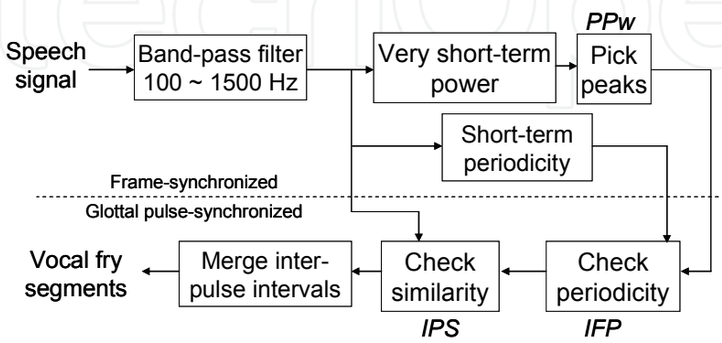


Fig. 5. Simplified block diagram of the vocal fry detection procedure.

3.3 Detection of pressed voice: H1'-A1'

Creaky voice (vocal fry) utterances may be pressed or lax. Lax creaky voices occur when the vocal folds are relaxed and are related to boredom or sadness (Gobl & Ní Cassaide, 2003). On the other hand, pressed creaky voices indicate strong attitudes/feelings of admiration or suffering (Sadanobu, 2004) in Japanese. Therefore, detection of pressed voices is necessary for PI disambiguation.

The production mechanism of pressed voice is not clearly explained yet, but it is thought that pressed voice has features similar with "tense voices" (Gordon & Ladefoged, 2001). A difference between "tense voice" and "lax voice" is reported to appear in the spectral tilt, since in tense voices, the glottal excitations become more impulse-like, and the higher frequency components are emphasized in relation to the fundamental frequency component. Acoustic parameters like H1-H2 and H1-A1 (Gordon & Ladefoged, 2001), and H1-A3 (Hanson, 1997) are proposed to reflect the effects of spectral tilt, where H1 is the amplitude power of the first harmonic (fundamental frequency), H2 is the amplitude power of the second harmonic, and A1 and A3 are the amplitude powers of the harmonic closest to the first and third formant, respectively.

However, in creaky or harsh voices, the irregularities in periodicity cause disturbances in the harmonic structure of their spectrum, so that it becomes difficult or unviable to extract harmonic components from the spectrum. In the present research, when periodicity is not detected, instead of H1, we use the maximum peak power of a low frequency band of 100 to 200 Hz (H1'). Also, as an automatic formant extraction is difficult, instead of A1, we use the maximum peak power in the frequency band of 200 to 1200 Hz (A1'), where the first formant is likely to appear. If periodicity is detected, H1' is equalized to H1. H1'-A1' values are estimated for each frame. Preliminary experiments indicate that pressed voice can be detected, when H1'-A1' is smaller than -15 dB.

3.4 Detection of aspiration noise: F1F3syn, A1-A3

Aspiration noise refers to turbulent noise caused by an air escape at the glottis, due to insufficient closure of the vocal folds during whispery and breathy phonations. Although there is a distinction between whispery and breathy voices from a physiological viewpoint (Laver, 1980), a categorical classification of voices in whispery or breathy is difficult in both acoustic and perceptual spaces (Kreiman & Gerratt, 2000). Further, aspiration noise can also occur along with harsh voices, composing the harsh whispery voices (Laver, 1980). In the present research, we use a measure of the degree of aspiration noise as indicative of such voice qualities.

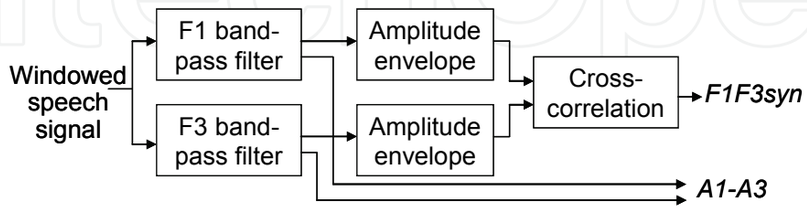


Fig. 6. Simplified block diagram of the acoustic parameters for aspiration noise detection.

The aspiration noise detection is based on the algorithm proposed in Ishi (2004), and its block diagram is shown in Fig. 6. The algorithm depends basically on two parameters.

- F1F3syn: synchronization measure between the amplitude envelopes of the signals in the first and third formant bands;
- A1-A3: difference (in dB) of the amplitudes of the signals in the first and third formant bands.

The main parameter, called F1F3syn, is a measure of synchronization (using a cross-correlation measure) between the amplitude envelopes of the signals obtained by filtering the input speech signal in two frequency bands, one around the first formant (F1) and another around the third formant (F3). This parameter is based on the fact that around the first formant, the harmonic components are usually stronger than the noisy component in modal phonation, while around the third formant, the noisy component becomes stronger than the harmonic components in whispery and breathy phonations (Stevens, 2000). Thus, when aspiration noise is absent, the amplitude envelopes of F1 and F3 bands are synchronized, and F1F3syn takes values close to 1, while if aspiration noise is present, the amplitude envelopes tend to be dissynchronized, and F1F3syn takes values closer to 0.

The second parameter, called A1-A3, is a measure of the difference (in dB) between the powers of F1 and F3 bands. This parameter is used to constrain the validity of the F1F3syn measure, when the power of F3 band is much lower than that of F1 band, so that aspiration noise could not be clearly perceived. Thus, when A1-A3 is big (i.e., the power of F1 band is much stronger than the power of F3 band), it is possible that the noisy components of F3 band are not perceived, and consequently, there is no sense to evaluate the F1F3syn measure.

The F1 band is set to 100 ~ 1500 Hz, while the F3 band is set to 1800 ~ 4500 Hz. The amplitude envelopes are obtained by taking the Hilbert envelope (Schroeder, 1999) of the signals filtered in each frequency band. Aspiration noise is detected for each frame, when F1F3syn is smaller than 0.4 and A1-A3 is smaller than 25 dB. These thresholds are based on the analysis results reported in Ishi (2004). More details about the evaluation of the method can be found in Ishi (2004).

3.5 Detection of harsh voice

A very simple procedure is adopted in the present research for detection of aperiodicity which is characteristic of harsh voices. The aperiodicity of harsh voices is here detected when neither periodicity nor vocal fry is detected. Note that vocal fry is usually aperiodic but does not sound harsh, so the aperiodicity in vocal fry segments has to be eliminated. Further, the initial and final 3 frames (30 ms) of each utterance are also eliminated, for avoiding the misdetection of aperiodicity due to effects of F0 disturbances at the onset and offset of the syllables.

Note that such simple procedure is valid, since we are evaluating only monosyllabic utterances and assuming that the voiced segments are known. Otherwise, the development of a more elaborated algorithm will be necessary for detecting harshness.

3.6 Evaluation of automatic detection of voice qualities

Fig. 7 shows a summary of the results for automatic detection of the voice qualities discussed in the previous sections.

The detection of creaky voice (or vocal fry) is evaluated by an index called VFR (Vocal Fry Rate), defined as the duration of the segment detected as vocal fry (VFdur) divided by the total duration of the utterance. Fig. 7 shows the results of detection of creaky segments, by using a criterion of VFR is larger than 0.1. We can note that all creaky segments are correctly detected (about 90% for c, c?), with only a few insertions (non c).

For evaluating pressed voice detection, an index called PVR (Pressed Voice Rate) is defined as the duration of the segment detected as pressed (PVdur), divided by the utterance duration. An utterance is detected as pressed, if PVR is larger than 0.1, and PVdur is larger than 100 ms, indicating that the segment has to be long enough to be perceived as pressed. 69 % of the pressed voice utterances were correctly identified in (p, pc, p?). Among them, most "e" utterances were correctly identified, while the detection failed in most of "un" utterances. This is probably because the nasal formant in "un" (around 100 to 300 Hz) increases the spectral power in the lower frequencies, consequently raising the H1'-A1' value. More robust acoustic features have to be investigated for detecting pressed voice in nasalized vowels.

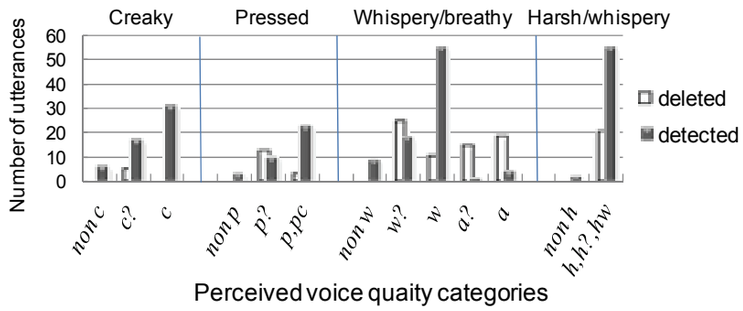


Fig. 7. Results of automatic detection of voice qualities, for each perceived category.

As in the previous voice qualities, an index called ANR (Aspiration Noise Rate) is defined as the duration of the segment detected as aspirated (ANdur), divided by the total duration of the utterance. Utterances containing aspiration noise are detected by using a criterion of ANR larger than 0.1. Most of the utterances where strong aspiration noise was perceived throughout the utterance (w) could be correctly detected (81%). However, for the utterances where aspiration noise was perceived in the syllable offsets (a? and a), most utterances could

not be detected by using ANR, as shown by the white bars in Fig. 8. This is because these syllable offset aspirations are usually unvoiced, and very short in duration. Other methods have to be investigated for the detection of the syllable offset aspirations.
Finally, regarding harsh and/or whispery voices, no clear distinction in functionality could be observed between harsh, harsh whispery and whispery voices (h, hw, w, a), as shown in Fig. 3. All these voice qualities are then described by an index called HWR (Harsh Whispery Rate). HWR is defined as the summation of HVdur (duration of the segment detected as harsh voice) and ANdur, divided by the utterance duration. 73 % of the utterances perceived as harsh and/or whispery (h,h?,hw) could be detected by using HWR > 0.1, and only a few insertion errors were obtained (non h), as shown in Fig. 7.

4. Discrimination of paralinguistic information based on intonation and voice quality-related prosodic features

In this section, we evaluate the contributions of intonation and voice quality-related prosodic features in "e"/"un" utterances, for discrimination of the PI items.

In 29 of the total of 370 utterances for evaluation, F0move could not be estimated due to missing F0 values. These missing values are due to non-modal phonations causing irregularities in the periodicity of the vocal folds. Fig. 8 shows a scatter plot of the intonation-related prosodic features (F0move vs. duration), excluding the utterances where F0move could not be obtained due to missing F0 values, and the ones where fall-rise intonation was detected.

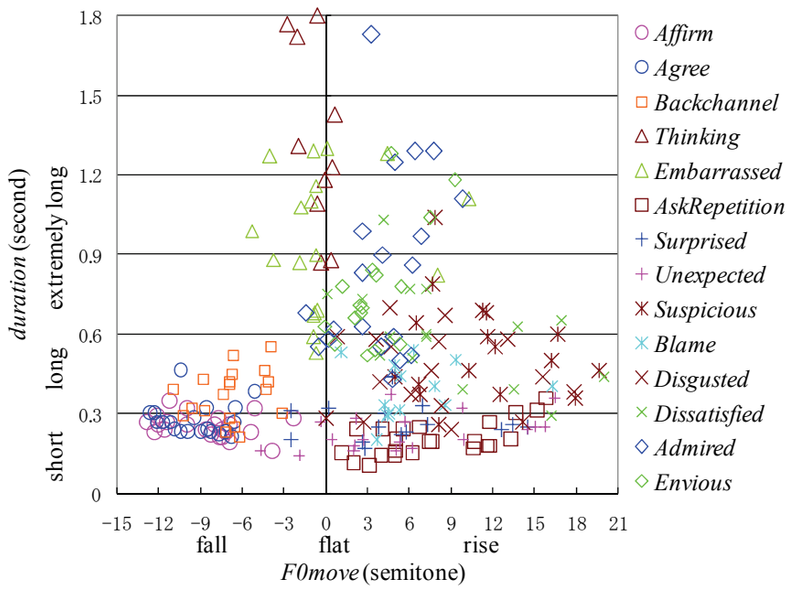


Fig. 8. Distribution of the intonation-related prosodic features (F0move vs. duration) for each PI.

Thresholds for F0move and duration were set, based on a preliminary evaluation of classification trees for discriminating the present PI items. A threshold of -3 semitones was set for F0move to discriminate falling tones (Fall), while a threshold of 1 semitone was set for rising tones (Rise). Utterances where F0move is between -3 and 1 semitone were considered as flat tones (Flat). The 29 utterances, where F0move could not be obtained, were

also treated as flat tones in the evaluation of automatic detection. Two thresholds were also set for duration. Utterances shorter than 0.36 seconds are called Short, while utterances with duration between 0.36 and 0.6 seconds are called Long. Utterances longer than 0.6 seconds are called extremely long (Ext.L).

Fig. 9 shows the distributions the prosodic categories (intonation and voice quality features) for each PI item. The discrimination of all PI items is difficult since many PI items share the same speaking styles. For example, there is no clear distinction in speaking style between Affirm and Agree, or between Surprised and Unexpected. The PI items which share similar speaking styles and which convey similar meanings in communication were then grouped (according to the perceptual evaluations in Section 2.2), for evaluating the automatic detection. Vertical bars in Fig. 9 separate the PI groups.

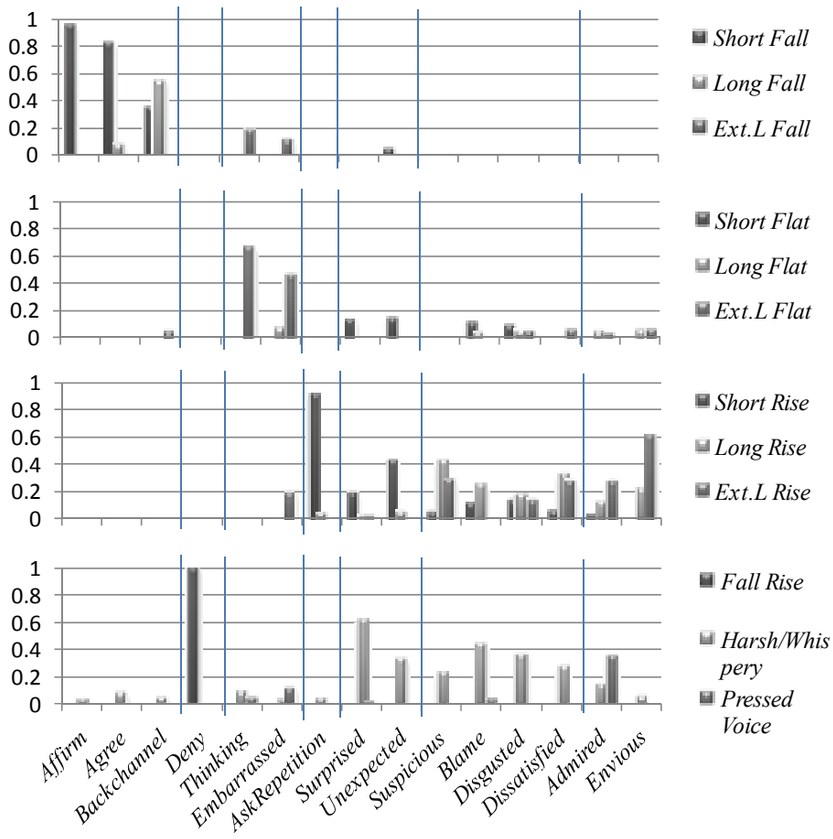


Fig. 9 Distribution of the prosodic categories for each PI item. Vertical bars separate PI groups.

Among the positive reactions, Affirm tends to be uttered by Short Fall intonation, while longer utterances (Long Fall) are more likely to appear in Backchannel. Extremely long fall or flat tones (Ext.L Fall, Ext.L Flat) were effective to identify Thinking. Note that the intonation-related prosodic features were effective to discriminate groups of PI items expressing some intentions or speech acts (Affirm/Agree/Backchannel, Deny, Thinking, and AskRepetition).

Short Rise tones can identify AskRepetition, Surprised and Unexpected, from the other PI items. Part of the Surprised/Unexpected utterances in Short Rise could be discriminated

from AskRepetition by the detection of harsh/whispery voice quality. However, many utterances in Surprised/Unexpected have similar speaking styles with AskRepetition. In these cases, context information would be necessary for their discrimination. The Fall-Rise tone detection was enough for the identification of Deny in the "un" utterances.

The big overlap of several PI items in the Rise tones shown in Fig. 8 resulted in lower discrimination between Surprised/Unexpected, Suspicious/Blame/Disgusted/Dissatisfied, and Admired/Envious, as shown in the second panel of Fig. 9. However, the use of voice quality features was effective to improve part of their detection rates. Note that the utterances where these non-modal voice qualities were detected pertain to the groups of PI items expressing strong emotions or attitudes.

Harsh and/or whispery voice detection (in the bottom panel of Fig. 9) was effective to disambiguate part of Surprised/Unexpected and AskRepetition sharing Short Rise tones, and Suspicious/Blame/Disgusted/Dissatisfied and Thinking/Embarrassed sharing Long and Extremely Long Flat tones.

Pressed voice detection was effective for identifying part of the utterances expressing admiration. The discrimination of Pressed utterances appearing in Disgusted and Embarrassed, from the ones in Admired, would need context information. However, it was observed that most utterances in Admired were "he", while most utterances in Disgusted and Embarrassed were "e", so that the detection of the aspirated consonant could be useful to discriminate part of the utterances.

Table 2 summarizes the detection rates without and with inclusion of voice quality features, for each PI group. Results indicate detection rates higher than 90 %, for Affirm/Agree/ Backchannel, Deny, and AskRepetition, regardless the use of voice quality features. For the three bottom PI groups (listed in Table 2) expressing strong emotions and attitudes, a significant improvement is obtained by the inclusion of voice quality features. However, the detection rates for Surprised/Unexpected and Suspicious/Blame/Disgusted/Dissatisfied were poor (41.9 % and 57.9 %). Improvements on the voice quality detection algorithms could still reduce part of these detection errors. However, most of the detection errors are thought to be due to the use of the same speaking style for different PI items, implying context dependency. Note that the confusions between PI items in Fig. 9 are basically the same as the ones obtained for the perceptual experiments in Section 2.2 (Fig. 2).

	Total	Detection rate (%) (without VQ)	Detection rate (%) (with VQ)	
Affirm/Agree/Backchannel	68	97.1	97.1	
Deny	12	100.0	100.0	
Thinking/Embarrassed	47	89.4	89.4	
AskRepetition	23	95.6	95.6	
Surprised/Unexpected	74	27.0	41.9	
Suspicious/Blame/Disgusted/Dissatisfied	88	38.6	57.9	83.6
Admired/Envious	58	39.7	63.8	
All PI items	370	57.3	69.2	86.2

Table 2. Detection rates of PI groups, without and with inclusion of voice quality (VQ) features.

The overall detection rate using simple thresholds for discrimination of the seven PI groups shown in Table 3 was 69.2 %, where 57.3 % was due to the only use of intonation-related prosodic features, while 11.9 % was due to the inclusion of voice quality parameters.
Finally, if the three PI groups Surprised/Unexpected, Suspicious/Blame/Disgusted/ Dissatisfied and Admired/Envious could be considered as a new group of PI items expressing strong emotions or attitudes, the detection rate of the new group would increase to 83.6 %, while the overall detection rate would increase to 86.2 %, as shown in the right-most column of Table 2. This is because most of the confusions in the acoustic space were among these three groups.

5. Conclusion

We proposed and evaluated intonation and voice quality-related prosodic features for automatic recognition of paralinguistic information (intentions, attitudes and emotions) in dialogue speech. We showed that intonation-based prosodic features were effective to discriminate paralinguistic information items expressing some intentions or speech acts, such as affirm, deny, thinking, and ask for repetition, while voice quality features were effective for identifying part of paralinguistic information items expressing some emotion or attitude, such as surprised, disgusted and admired. Among the voice qualities, the detection of pressed voices were useful to identify disgusted or embarrassed (for "e", "un"), and admiration (for "he"), while the detection of harsh/whispery voices were useful to identify surprised/unexpected or suspicious/disgusted/blame/dissatisfied.
Improvements in the detection of voice qualities (harshness, pressed voice in nasalized voices, and syllable offset aspiration noise) can still improve the detection rate of paralinguistic information items expressing emotions/attitudes.
Future works will involve improvement of voice quality detection, investigations about how to deal with context information, and evaluation in a human-robot interaction scenario.

6. Acknowledgements

This research was partly supported by the Ministry of Internal Affairs and Communications and the Ministry of Education, Culture, Sports, Science and Technology-Japan. The author thanks Hiroshi Ishiguro (Osaka University), Ken-Ichi Sakakibara (NTT) and Parham Mokhtari (ATR) for advice and motivating discussions.

7. References

Campbell, N., & Erickson, D. (2004). What do people hear? A study of the perception of non-verbal affective information in conversational speech. *Journal of the Phonetic Society of Japan*, 8(1), 9-28.

Campbell, N., & Mokhtari, P. (2003). Voice quality, the 4th prosodic dimension. In *Proceedings of 15th International Congress of Phonetic Sciences (ICPhS2003),* Barcelona, (pp. 2417-2420).

Dang, J., Honda, K. (1966). Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.*, 101(1), 456-465.

Erickson, D. (2005). Expressive speech: production, perception and application to speech synthesis. *Acoust. Sci. & Tech*., *26*(4), 317-325.

Fernandez, R., & Picard, R.W. (2005). Classical and novel discriminant features for affect recognition from speech. In *Proceedings of Interspeech 2005*, Lisbon, Portugal (pp. 473-476).

Fujie, S., Ejiri, Y., Matsusaka, Y., Kikuchi, H., & Kobayashi, T. (2003) Recognition of paralinguistic information and its application to spoken dialogue system. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* (*ASRU '03*), St. Thomas, U.S. (pp. 231-236).

Fujimoto, M., & Maekawa, K. (2003) Variation of phonation types due to paralinguistic information: An analysis of high-speed video images. In *Proceedings of 15th International Congress of Phonetic Sciences (ICPhS2003)*, Barcelona, (pp. 2401-2404).

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, 189-212.

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *J. of Phonetics*, 29, 383-406.

Hanson, H. (1997). Glottal characteristics of female speakers: Acoustic correlates. *J. Acoustic Society of America*, 101, 466-481.

Hayashi, Y. (1999). Recognition of vocal expression of emotions in Japanese: using the interjection "eh". In *Proceedings of ICPhS 99*, San Francisco, USA (pp. 2355-2359).

Hess, W. (1983). Pitch Determination of Speech Signals. Vol. 3 of *Springer Series of Information Sciences*, Berlin, Heidelberg, New York: Springer-Verlag.

Imagawa, H., Sakakibara, K., Tayama, N., & Niimi, S. (2003). The effect of the hypopharyngeal and supra-glottic shapes for the singing voice. In *Proceedings of the Stockholm Music Acoustics Conference (SMAC 2003)*, II, (pp. 471-474).

Ishi, C.T. (2004). A new acoustic measure for aspiration noise detection. In *Proceedings of Interspeech 2004-ICSLP*, Jeju, Korea (pp. 941-944).

Ishi, C.T. (2005). Perceptually-related F0 parameters for automatic classification of phrase final tones. *IEICE Trans. Inf. & Syst.*, E88-D(3), 481-488.

Ishi, C.T., Ishiguro, H., & Hagita, N. (2005). Proposal of acoustic measures for automatic detection of vocal fry. In *Proceedings of Interspeech 2005*, Lisbon, Portugal (pp. 481-484).

Ito, M. (2004). Politeness and voice quality – The alternative method to measure aspiration noise. In *Proceedings of Speech Prosody 2004*, Nara, Japan (pp. 213-216).

JST/CREST ESP Project homepage, http://feast.atr.jp/esp/esp-web/

Kasuya, H., Yoshizawa, M., & Maekawa, K. (2000). Roles of voice source dynamics as a conveyer of paralinguistic features. In *Proceedings of International Conference on Spoken Language Processing (ICSLP2000)*, Beijing, (pp.345-348).

Kitamura, T., Honda, K., Takemoto, H. (2005). Individual variation of the hypopharyngeal cavities and its acoustic effects. Acoust. Sci. & Tech 26(1), 16-26.

Klasmeyer, G., & Sendlmeier, W.F. (2000). Voice and Emotional States. In R.D. Kent & M.J. Ball (Eds.), *Voice Quality Measurement*, San Diego: Singular Thomson Learning, 339-358.

Kreiman, J., & Gerratt, B. (2000). Measuring vocal quality. In R.D. Kent & M.J. Ball (Eds.), *Voice Quality Measurement*, San Diego: Singular Thomson Learning, 73-102.

Laver, J. (1980). Phonatory settings. In *The phonetic description of voice quality*. Cambridge: Cambridge University Press, 93-135.

Maekawa, K. (2004). Production and perception of 'Paralinguistic' information. In *Proceedings of Speech Prosody 2004*, Nara, Japan (pp. 367-374).

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. In *Proceedings of Interspeech 2006*, Pittsburgh, USA (pp. 809-812).

Nwe, T.L., Foo, S.W., & De Silva, L.C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication* 41, 603-623.

Sadanobu, T. (2004). A natural history of Japanese pressed voice. *J. of Phonetic Society of Japan*, 8(1), 29-44.

Schroeder, M.R. (1999). Hilbert envelope and instantaneous frequency. In *Computer speech – Recognition, compression, synthesis*, Berlin: Springer, 174-177.

Schuller, B., Muller, R., Lang, M., & Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proceedings of Interspeech 2005*, Lisbon, Portugal (pp. 805-808).

Stevens, K. (2000). Turbulence noise at the glottis during breathy and modal voicing. In *Acoustic Phonetics*. Cambridge: The MIT Press, 445-450.

Voice quality sample homepage, http://www.irc.atr.jp/~carlos/voicequality/



Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9

Hard cover, 550 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Carlos T. Ishi (2008). Recognition of Paralinguistic Information using Prosodic Features Related to Intonation and Voice Quality, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from: http://www.intechopen.com/books/speech_recognition/recognition_of_paralinguistic_information_using_prosodic_features_related_to_intonation_and_voice_qu

INTECH
open science | open minds

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2008 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen